DOCUMENT RESUME

ED 233 057 TM 830 504

AUTHOR Phillips, S. E.; Anderson, A. E.
TITLE LOGTRUE: A Computer Program for Test Equating with Item Response Theory.
PUB DATE Apr 83
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Programs; *Equated Scores; Group Testing; *Latent Trait Theory; Raw Scores; *Scaling; Test Interpretation; Test Items
IDENTIFIERS LOGIST Computer Program; *LOGTRUE Computer Program

ABSTRACT

The LOGTRUE program can be used to obtain a scale of equated raw scores for two tests with parameter estimates on a common item response theory scale. The program derives its name from the method of logistic true score equating described by Lord (1980). The method can be applied to two tests with overlapping items administered to different groups of examinees (anchor test approach) or to two non-overlapping tests administered to the same group of examinees (scaling test approach). The method requires equal ability distributions for the examinees completing the two tests being equated, or that parameter estimates for the items in the two tests be transformed to a common scale based on a subset of overlapping items. Equating then proceeds. LOGTRUE's input data are the output of the LOGIST or EQUATOR computer programs. Difficulty estimates and the effects of difficulty parameter outliers are examined. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233057

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

LOGTRUE:

A COMPUTER PROGRAM FOR TEST EQUATING WITH ITEM RESPONSE THEORY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. E. Phillips
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

S. E. Phillips

A. E. Anderson

Michigan State University

TM 830 509

A paper presented at the AERA Annual Meeting,
Montreal 1983

Printed in the USA

LOGTRUE: A Computer Program For Test Equating With Item Response Theory

The program described in this document can be used to obtain a scale of equated raw scores for two tests with parameter estimates on a common item response theory scale. The program derives its name, LOGTRUE, from the method of logistic true score equating described by Lord (1980).

A variety of approaches have been used by test publishers to derive a common score scale through the horizontal or vertical equating of test forms. In their most basic form, these include the complete, anchor, and scaling test methods.

The complete test method requires the administration of complete tests to samples of students at every grade level at each time of year for which norms are to be provided. Due to the large number of test administrations and time required, this method is used infrequently. The scaling test method relies on the administration of a special scaling test spanning a wide range of difficulty and content. Results of the scaling test at each grade level are linked to the complete test form designed for that grade by administering both tests to the same group of students. In the anchor test method, student performance in adjacent grades is linked by the administration of overlapping tests. Mittman (1958) recommends the scaling test approach over the complete and anchor test methods for scaling achievement tests. For other applications, the anchor test method may be more suitable.

The logistic true score equating method can be applied to two tests with overlapping items administered to different groups of examinees (anchor test approach) or to two non-overlapping tests administered to the same group of

examinees (scaling test method). Before true score equating is applied to an anchor test situation, the item parameters for all test items must be converted to the same scale. One method for doing so is Haebara's (1980) EQUATOR program which utilizes a weighted least squares method. Item parameters in the scaling test method will already be on the same scale because the two sets of responses are based on a common ability distribution (the same examinees).

Related Software

LOGTRUE has as its input data the output of the LOGIST (Wood, Wingersky, and Lord, 1976) computer program or the EQUATOR program (Haebara, 1980). At present, the program is not designed to handle the output of BICAL (Wright and Mead, 1976).

The LOGTRUE program was written for a CDC Cyber system and runs under the SCOPE/HUSTLER operating system. A similar program developed at Educational Testing Service on an IBM system includes the functions of the LOGTRUE and EQUATOR programs (GEEQUATE - Stocking, 1982).

Methodology

The method of true score equating in item response theory as described by Lord (1980) requires equal ability distributions for the examinees completing the two tests being equated or that parameter estimates for the items in the two tests have been transformed to a common scale based on a subset of overlapping items. The equating then proceeds as follows.

Assume two tests ($i = 1, 2$) with j items having item parameters a_{ij}, b_{ij}, and c_{ij} (discrimination, difficulty, and lower asymptote, respectively). Let n_i be the number of items in test i, $P(\theta)$ be the expected raw score for an ability θ, and D an arbitrary scaling constant equal to 1.7. Then the following equations define the two test characteristic curves:

$$P_1(\theta) = c_{1j} + (1-c_{1j}) \frac{\exp(Da_{1j}(\theta-b_{1j}))}{1+\exp(Da_{1j}(\theta-b_{1j}))} \quad (1);$$

$$P_2(\theta) = c_{2j} + (1-c_{2j}) \frac{\exp(Da_{2j}(\theta-b_{2j}))}{1+\exp(Da_{2j}(\theta-b_{2j}))} \quad (2).$$

To equate test two to test one, the procedure is as follows. First, the θ (ability) value associated with each possible integer raw score is estimated for test two. These ability values corresponding to integer raw scores are computed from equation two using an iterative estimation algorithm. Then these estimated abilities (θ's) for each integer raw score on test two are substituted into equation one. The resulting $P(\theta)$ values are the equated fractional raw scores for test one.

In solving for the argument of a function when its value is known (i.e., finding θ given $P(\theta)$), the obvious technique (if the function is simple enough) is to derive the inverse function and "plug in" Y. That method is not likely to be feasible, however, if the function is a composite test characteristic curve with 50 items and 150 parameters. The most common algorithm used in such a situation is the "binary chop" where the "answer" X is trapped between a low and a high bound. A new guess is tried at the midpoint (low + high)/2 and plugged into the function. If the result is less than the target Y, the desired X must be in the upper half of the

interval; otherwise X must be in the lower half of the interval. This process is guaranteed to converge properly only if the function is continuous and monotonic. Fortunately, both are true for the sensitive region of the characteristic curve of every test with positively discriminating items, and is true of the entire curve for most tests.

It can be shown that the "binary-chop" is the optimal algorithm if the desired X has a uniform probability distribution over the interval (low, high). However we know several more things: f(low), f(high) and that f is almost linear over short intervals. Therefore a "proportional-chop" algorithm is used which produces quadratic convergence instead of linear. The problem with this technique is that if the desired X is close to that end of interval where the slope is flatter, the guess will always be "outside" the desired X, and the interval will hardly shrink at all. The "fix" is to move the guess toward the midpoint a small amount. This may be done by taking a weighted average of the guess and the midpoint. Then the guess will "move over" more if it is close to an endpoint of the interval. If the weight given the midpoint depends on the size of the interval then the algorithm performs as well as binary-chop on large intervals and much better on very small intervals where the function is more linear.

Item response theory models other than the three-parameter model can be used with the LOGTRUE equating program by setting parameter values to the appropriate constants. For example, for a two-parameter model, all c_{ij} would be set to zero.

If a three-parameter model is utilized, the true score equating procedure described in this section can yield raw scores only as small as the sum of the lower assumptotes (Σc_{ij}) for the given test. This can be demonstrated as follows:

$$\text{Let } Y = Da_j\ (\theta - b_j)$$

$$\text{Then } P(\theta) = \Sigma c_j + \Sigma\ \frac{(1-c_j)\ \exp\ (Y)}{1 + \exp\ (Y)}$$

$$\lim_{\theta \longrightarrow -\infty} P(\theta) = \Sigma c_j + \Sigma\ \frac{(1-c_j)\ \exp\ (-\infty)}{1 + \exp\ (-\infty)}$$

$$\text{(where } \exp\ (-\infty) = 1 \,/\, \exp\ (\infty) = 1/\infty = 0)$$

$$= \Sigma c_j + 0$$

$$= \Sigma c_j .$$

Lord (1980) suggests the following procedure for determining equated raw scores for values less than the sum of the c_{ij} parameters. Consider a hypothetical group of examinees with $\theta = -\infty$. This group will have a mean test score of Σc_{ij} with variance $\Sigma c_{ij}\ (1 - c_{ij})$. A linear equating function is constructed based on this group of examinees. Letting X = test one and Y = test two, the equating function x (y) for converting test two scores to the scale of test one scores below the chance level is:

$$x\ (y)\quad \frac{\sqrt{\Sigma c_{1j}\ (1-c_{1j})}}{\sqrt{\Sigma c_{2j}\ (1-c_{2j})}}\quad (y - \Sigma c_{2j}) + \Sigma c_{1j}.$$

Use of this equating function below the chance level and true score equating above the chance level produces a continuous equating relationship.

Outlines in Difficulty Estimates

If the proportion of examinees answering an item correctly is very large or very small, the difficulty parameter of the item will have a large standard error and its estimate will be unstable (Lord, 1980). Together with an item characteric curve with a very flat slope, this may result in difficulty estimates with very large absolute values (e.g., -128.6 or 58.7). Generally, such large positive or negative b values will be rare and will fall substantially outside of the range of b parameters estimated for the other items in the test. Since a very slight change in the slope of the item characteristic curve of such an item can result in a large jump in the b value (e.g., 128.6 to 153.7), it appears that values outside the usual range are not subject to the same interpretation as those within the range of the other test items.

In an equating situation, one possible way of reducing the effect of items with difficulty parameters which are not well estimated is to set a maximum and minimum value for all difficulty parameters. Any b which exceeds these limits in absolute value would be replaced by the appropriate upper or lower limit.

To determine the magnitude of the equating effects of such a procedure, it was applied to test data from a previous equating study (Phillips, 1982). LOGIST was used to estimate item parameters using a three parameter model for achievement test data for mathematics and reading subtests in grades four and eight. The equating was based on the scaling test method and is described more fully in the original study (Phillips, 1982).

For purposes of comparison four different cutoff limits were established for the difficulty parameter including ± 4, ± 9, ± 15, ± 25, ± 50, and ± 100. In

each case, prior to equating, all difficulty values for the test items which exceeded the upper cutoff limit or were less than the lower cutoff limit were replaced in the equating by the appropriate cutoff value. The frequency of changed b values for each set of cutoff points by grade and subject is listed in Table 1. Few items in the regular battery tests exceed the cutoff points. The number of items with difficulties outside the cutoff points ranged from 0 to 17 and were similar for 15, 25, and 50.

- -

Insert Table 1 about here

- -

The test equating followed the logistic true score equating procedure outlined by Lord (1980). For comparison, the equating was also completed for each subject/grade combination using the original difficulty values estimated by the LOGIST program. The equating results using the different cutoff points for the difficulty parameters were then compared to the results when no limits were placed on the difficulty values.

The average differences in raw score equated values are presented in Table 2. As expected the differences decrease as the absolute value of the cutoff points increases. The differences by raw score were very consistent with the averages in both magnitude and direction for all subject/grade combinations and so have not been reported directly in this study.

- -

Insert Table 2 about here

- -

The averages in Table 2 range from -2.3 to 1.6 raw score points. Use of cutoff values appeared to lower the equated raw scores in grade four but to increase them in grade eight. Reading grade eight is unusual because

no items had difficulty values falling outside cutoff points of nine or more.

The difficulty parameter outliers do appear to affect the equating results for this type of data. More empirical research is needed to determine whether outliers should be retained in the analysis or whether better equating results can be obtained if cutoffs are established. For the equating program described in this document the authors have made use of a maximum absolute difficulty an option of the program.

Table 1

Number of Items with Absolute Difficulty Parameters Excceeding the Cutoff

	Cutoff (Regular Test/Scaling Test)					
	4	9	15	25	50	100
Math Grade 4	3/13	1/11	1/8	1/7	1/6	1/4
Math Grade 8	1/17	0/5	0/2	0/2	0/2	0/1
Read Grade 4	0/8	0/5	0/5	0/5	0/5	0/4
Read Grade 8	0/5	0/0	0/0	0/0	0/0	0/0

Table 2

Average Difference in Equated

Raw Scores by Cutoff

Cutoff Points for b-Values

	±4	±9	±15	±25	±50	±100
Math Grade 4	-2.31	-1.47	-1.27	-1.02	-.72	-.18
Math Grade 8	.55	.68	.50	.38	.20	.06
Read Grade 4	-1.55	-1.46	-1.35	-1.20	-.82	-.20
Read Grade 8	.29	0	0	0	0	0

References

Haebara, T. Equating Logistic Ability Scales by a Weighted Least Squares Method. Iowa Testing Programs Occasional Papers, The University of Iowa, Number 27, August, 1980.

Lord, F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey, Lawrence Erlbaum Associates, 1980.

Mittman, A. An Empirical Study of Methods of Scaling Achievement Tests at the Elementary Grade Level. Unpublished doctoral dissertation, University of Iowa, Iowa City, Iowa, 1958.

Phillips, S.E. Comparison of Equipercentile and Latent Trait Equating When the Scaling Test Method is Applied to a Multilevel Achievement Battery. 1982 (in press).

Stocking, M.L. Documentation for General Equating Program. Educational Testing Service, Princeton, New Jersey, 1982.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters. Research Memorandum, 76-6. Educational Testing Service, Princeton, New Jersey 1976.

Wright, B.D. and Mead, R.J. BICAL: Calibrating Items and Scales with the Rasch Model. Research Memorandum No. 23, University of Chicago Statistical Laboratory, Department of Education, 1976.